UNIT-2 Data Pre-Processing

Data Preprocessing : An Overview

- → Data Quality
- → major tasks in Data Preprocessing
  - → Data Cleaning
  - → Data Integration
  - → Data Reduction
  - → Data Transformation & data discretization

Data pre-processing is the process of preparing the data & making it suitable for ML model.

- → Importance of data pre-processing
  (Data in raw format contains noises, missing value, not in useable format)
- → Data Quality Measures
  - → Accuracy
  - → Completeness
  - → Consistency
  - → Timliness
  - → Bellevability
  - → Interpretability

Major Tasks in Data Preprocessing

→ Data Collection
- → Data Cleaning
- → Data Integration
- → Data reduction
- → Data transformation & data discretization

① **Data Cleaning** :- filling missing value, smooth out noise, identify outliers.

② **Data integrity** :- Data integration is a technique to merges data from multiple sources into a Coherent data store.

③ **Data reduction** :- Data reduction is a technique to reduce the data size by aggregating, eliminating redundant features

④ **Data transformation** :- Data are transformed or consolidated into forms appropriate for ML model training.

↳ **data discretization** :- Transforms the numeric data by mapping values to interval or Concept level.

Data Collection

Gathering data from various sources such as

→ Social media
→ databases
→ External repositories
→ IOT devices
→ Multimedia data

→ Different data collection methods

- → Primary data Collection
- → Secondary data Collection

Primary data collection

- → Surveys & Questionnaries
- → Interviews
- → Observations
- → Experiments
- → focus Groups

Secondary data collection

- → Published Sources
- → Online Databases
- → Government & Institutional records
- → Publicly available data
- → Past research studies

→ Data Collection Tools

- → Social media listening Tools
- → Web Analytics tools
- → Data Logging Devices
- → Mobile Data Coll^{n} apps
- → IOT devices

Data Augmentation (A method of data Coll^{n})

(Expand the Size of Existing dataset without gathering more data)

- Methods
- Rotating the Original image
- Crop the original Image differently
- Altering the light conditions
- Ramdom Cropping and Padding
- Scaling & Zooming
- Shearing & Perspective transform
- Colour Segmentation
- Gaussian noise

- Types of Data Augmentation
  - Real data augmentation approches
    - Sensor noise
    - Occlusion
    - Weather
    - Time Series
    - Label Smoothing
  - Synthetic data augmentation approches
    - Image synthesis
    - Text generation
    - Oversampling and Undersampling
    - Data Interpolation & Extrapolation
    - feature Perturbation

- Challenges faced by Data augmentation
  - Data Security & Privacy
  - Maintaining label Integrity
  - May Increase Size of training dataset

Data Labelling

→ A way of Identifing the raw data & adding suitable labels or tags to data.

→ Specify what data is all about

→ Labelled Data v/s Unlabelled data

- Labelled data:
  - → has predefined tags
  - → Supervised ML
- Unlabelled data:
  - → no such tags or labels
  - → Unsupervised ML

→ Tools for Data labelling

- → Super annotate
- → Encord
- → Kili
- → Appen
- → Dataloop
- → labelles
- → Scale Rapid
- → Argo hub
- → Trainingdata.io

→ Approaches to Data labelling

- → Internal data labelling
- → Synthetic labelling
- → Programmatic labelling
- → Outsourcing
- → Crowdsourcing

Benfits of Data labelling

- Precise Predictions
- Better Data Usability

Challanges of Data labelling

- Costly & time consuming
- Possibilities of Human-errors

Data Cleaning

(Data cleaning is the process of correcting or removing inacurates improperty formated, duplicated or deal with missing values from dataset)

In general data cleaning lowers errors, improves the quality of the data

- Steps for Cleaning data
  - Remove duplicate or irrelevant observations
  - fix Structural errors
  - filter unwanted Outliers
  - Handle missing data
  - Validate & QA
- Techniques for Data cleaning
  - Ignore the tuples
  - fill in the missing value
  - Binning method
  - Regression
  - Clustering

→ Process of Data Cleaning

- → Monitoring the errors
- → Standardize the minning process
- → Validate data accuracy
- → Scrub for duplicate data
- → Research on data
- → Communicate with the team.

→ Tools for Data Cleaning

- → OpenRefine
- → Trifacta Wrangler
- → Drake
- → Data Ladder
- → Cloudingo
- → Reifier (any many more)

→ Benifits of Data cleaning

- → Removes inaccuracies
- → Capacity to map many functions
- → Monitoring Mistakes
- → Makes decisions more quickly.
- → Enhance the Efficiency

Data Integration

(Process of Merging data from several disparate sources)

- → Characteristics (G, S, M)
  - G → global Schema
  - S → heterogeneous source of schema
  - M → mapping b/w sources & global Schema
- → Data Integration Approaches
  - → Tight Coupling
  - → low Coupling
- → Issues in Data Integration
  - → Entity Identification Problem
  - → Redundancy & Correlation Analysis
  - → Tuple Duplication
  - → Data Warfare Detection & backbone
- → Data Integration Techniques
  - → Manual Integration
  - → Middleware Integration
  - → Application- based Integration
  - → Uniform access Integration
  - → Data warehousing
- → Integration Tools
  - → On- promise data integration tool
  - → open- source data integration tool
  - → cloud- based data integration tool

Data Reduction

(Procers of reducing the volume of Original data & represents it in a much Smaller volume)

- -> By reducing the data, the efficiency of data analysis process is improved, which produces the same analytical results.
- -> Importance of data reduction
  - -> Aims to define it in more compact form.
  - -> In terms of rows & colms.
- -> Techniques of Data Reduction
  - -> Dimensionality Reduction
  - -> Numerosity Reduction
  - -> Data Cube Aggregation
  - -> Data Compression

I -> Dimensionality Reduction

(Eliminates attributes from data set by reducing vol. of original data)

- -> A way of converting higher dimensions into lesser dimensions in a dataset.
- -> Methods of Dimensionality Reduction
  - -> Feature Selection
  - -> feature Extraction
  - -> Wavelet Transform
  - -> Attribute Subset Selection

Dimensionality Reduction

Feature Selection
- → Filter method
- → Wrapper method
- → Embedded method

Feature Extraction
- → Principal Component Analysis (PCA)
- → Factor Analysis
- → Singular Value decomposition

→ PCA Principal Component Analysis
- → Karl, Pearson, 1901
- → High dimensions → lower dimensions
- → Preserving most important patterns & relation between Variables

n - Total no. of dimensions in a dataset
k - independent features

$n > k$

→ Factor Analysis
(There will not be any Outliers, reduce a large no. of Variables into fewer no. of factors)

→ Singular value decomposition
(helps us to simplify the data, finds most important patterns in the data & focus on them.)

Wavelet Transformation

A data vector A is transformed into a numerically different data vector A' such that both A & A' vectors are of same length

Attribute Subset Selection

Reduces the volume of data by eliminating redundant & irrelevant attributes

(II) Numerosity Reduction

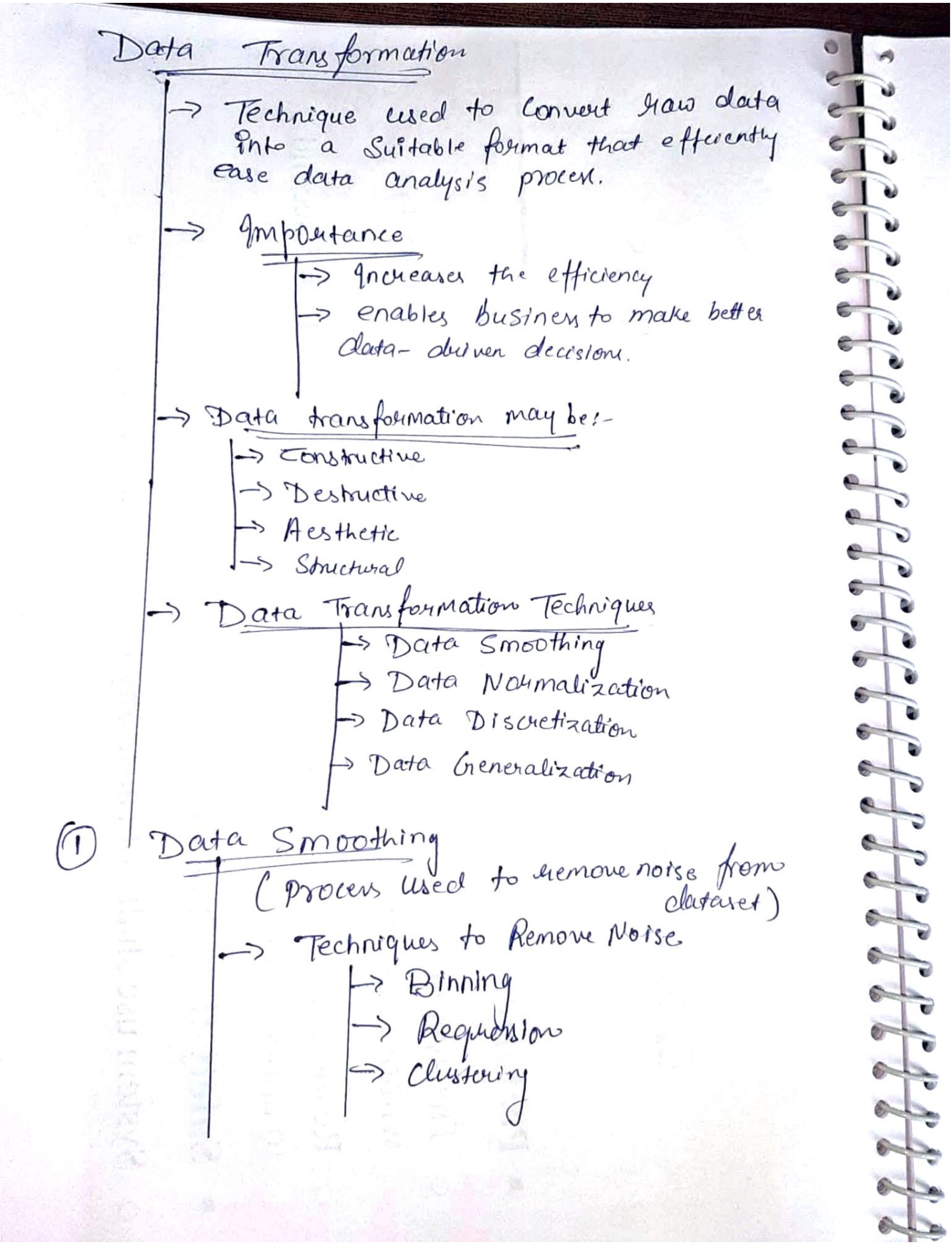
- → Reduces the data volume by chossing alternative smaller forms of data representation
- → Types of numerosity reduction
  - → Parametric numerosity reduction
  - → Non-Parametric numerosity reduction

(III) Data Cube Aggregation

- → This technique is used to aggregate data in a simpler form
- → It is a multidimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction

(IV) Data Compression

- → Employs modification, encoding & converting the structure of data in a way that consumes less space.
- → Types of Data Compression
  - → Lossless data compression
  - → Lossy data compression



Data Transformation

- Technique used to Convert raw data into a Suitable format that efficently ease data analysis's process.
- Importance
  - Increases the efficiency
  - enables business to make better data- driven decisions.
- Data transformation may be:-
  - Constructive
  - Destructive
  - Aesthetic
  - Structural
- Data Transformation Techniques
  - Data Smoothing
  - Data Normalization
  - Data Discretization
  - Data Generalization

(1) Data Smoothing

(Process used to remove noise from dataset)

- Techniques to Remove Noise
  - Binning
  - Regression
  - Clustering

② Data Normalization

(refers to Scale the data values to a much Smaller range)

- → Methods to Normalize the data
  - → Min-max normalization
  - → Z-Score normalization
  - → Decimal Scaling

③ Data Discretization

(process of converting continuous data into a set of data intervals.)

- → Classification of data discretization
  - → Supervised discretization
  - → Unsupervised discretization

④ Data Generalization

(Converts low-level data attributes to high-level data attributes)

- → Divided into two approches
  - → Data Cube process (OLAP)
  - → Attribute-oriented Induction (AOI)

→ Data Transformation process

(Entire process for transforming data is known as ETL)

- → Data discovery
- → Data Mapping
- → Data Extraction
- → Code Generation & Execution
- → Review
- → Sending.

Advantages of Data Transformation

- → Better Organization
- → Improved data quality
- → Perform faster Queries
- → Better data Management
- → More use out of Data

Disadvantages of Data Transformation

- → Can be Expensive
- → Can be resource-intensive
- → Lack of Expertise can introduce problem

Tools for Data Transformation

- → Scripting
- → On-premises ETL tools
- → Cloud-Based ETL tools